Environmental Science & Technology

Article

pubs.acs.org/est

Development, Evaluation, and Comparison of Land Use Regression Modeling Methods to Estimate Residential Exposure to Nitrogen Dioxide in a Cohort Study

Jonathan Gillespie,[†] Iain J. Beverland,[*,†] Scott Hamilton,[‡] and Sandosh Padmanabhan[§]

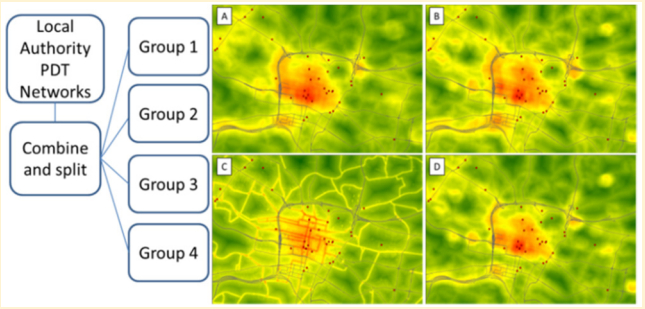
[†]Department of Civil and Environmental Engineering, University of Strathclyde, James Weir Building, 75 Montrose Street, Glasgow, G1 1XJ, U.K.

[‡]Ricardo Energy and Environment, 18 Blythswood Square, Glasgow G2 4BG, U.K.

[§]University of Glasgow, Institute of Cardiovascular and Medical Sciences, British Heart Foundation Glasgow Cardiovascular Research Centre, 126 University Place, Glasgow G12 8TA

Supporting Information

ABSTRACT: We used a network of 135 NO_2 passive diffusion tube sites to develop land use regression (LUR) models in a UK conurbation. Network sites were divided into four groups (32−35 sites per group) and models developed using combinations of 1−3 groups of "training" sites to evaluate how the number of training sites influenced model performance and residential NO_2 exposure estimates for a cohort of 13 679 participants. All models explained moderate to high variance in training and independent "hold-out" data (Training adj. R^2: 62−89%; Hold-out R^2: 44−85%). Average hold-out R^2 increased by 9.5%, while average training adj. R^2 decreased by 7.2% when the number of training groups was increased from 1 to 3. Exposure estimate precision improved with increasing number of training sites (median intralocation relative standard deviations of 19.2, 10.3, and 7.7% for 1-group, 2-group and 3-group models respectively). Independent 1-group models gave highly variable exposure estimates suggesting that variations in LUR sampling networks with relatively low numbers of sites (≤35) may substantially alter exposure estimates. Collectively, our analyses suggest that use of more than 60 training sites has quantifiable benefits in epidemiological application of LUR models.



1. INTRODUCTION

Traffic related air pollution (TRAP) is associated with premature mortality and adverse health outcomes.[1] However, the quantification and interpretation of such associations are made less certain by challenges of reliably estimating human exposure to air pollutants across extended spatial areas and temporal periods. The majority of exposure estimates in large cohort studies are based on modeled pollutant concentrations to address practical problems that prevent direct exposure assignment through monitoring of pollution concentrations in sufficiently numerous locations over sufficiently long time periods.

Land use regression (LUR) modeling is an extensively used exposure estimate approach that uses geographical information systems (GIS) and statistical analyses to quantify multivariate relationships between geographic features and measured TRAP concentrations.[2] LUR is relatively simple and cost-effective to implement, and can be applied to different atmospheric pollutants measured over large geographical areas[3] and to quantify associations between air pollution concentrations and meteorological,[4−6] temporal,[7,8] and building geometry information.[9,10]

A practical limitation is the requirement to conduct extensive monitoring prior to LUR model development, with studies typically monitoring at least 20 sites, two or more times, with subsequent extrapolation to estimate annual average exposures.[3,11,12] Limited numbers of monitoring sites combined with a high number of variables increases the risk of "overfitting" LUR models. To maximize the number of sites used in model development it is common to use all available data as "training sites" in model development retaining few, if any, independent "hold-out" sites for model evaluation. Model performance, quantified by explained variance of concentrations at the training sites used for model development, has been observed to decrease as the number of training sites is increased, while performance on quantification of variance of hold-out observations has been observed to increase as the number of training sites increases.[13−15] Leave one out cross validation (LOOCV) is often used to evaluate LUR model

Received: April 26, 2016
Revised: September 1, 2016
Accepted: September 12, 2016
Published: September 12, 2016

DOI: 10.1021/acs.est.6b02089
Environ. Sci. Technol. 2016, 50, 11085−11093

performance; however this approach may overestimate model performance.[16] A possible strategy to increase the number of monitoring sites is to use local authority managed monitoring networks, either alone or in combination with LUR "purpose designed" networks. In the United Kingdom local authorities operate passive diffusion tube (PDT) networks as part of statutory local air quality management activities, providing observations which can be used for LUR development and evaluation. Monitoring in local authority networks is conducted throughout the year, and hence provides annual mean estimates without extrapolation, although there is potential for "preferential sampling" through relatively high numbers of roadside monitoring locations which may bias model estimates.[17] A review in 2008 by Beelen et al.[3] identified 25 LUR studies, and of these, seven made use of data routinely collected by local authorities. Subsequently published research has mostly used purpose designed networks in LUR model development, although there are some examples employing routinely collected data.[18−22]

Our study developed LUR models for the Greater Glasgow conurbation using data routinely collected through local authority PDT networks. We examined the effect of randomized site selection on model development, and the effect of increasing the number of training sites. We used the LUR models to estimate residential exposure of 13 679 cohort participants at 9631 separate postcode address locations and examined the extent to which estimated NO_2 concentrations differed at each address location between LUR models developed from different groups of training sites to quantify the precision of exposure estimates. Our analyses enabled direct comparison of exposure estimates produced by LUR models developed from independent subsets covering the same spatial and temporal extent of the full network.

2. MATERIALS AND METHODS

2.1. Study Area and Monitoring Sites. The study focused on the Glasgow conurbation (55.865°, −4.260°; population approximately 1.2 million) in the west of Scotland. The conurbation includes local authority areas of Glasgow City, Renfrewshire, East Renfrewshire and East Dumbartonshire, whose combined PDT networks included a total of 218 sites during the decade preceding 2016.

Monthly PDT NO_2 observations obtained from local authorities[23] were converted to annual averages for each site, excluding sites with <75% annual data collection.[21,24] We selected 2007 for analyses to maximize the number of available monitoring sites (n = 135). Local authority PDT networks are often designed primarily to estimate compliance with national and international air quality objectives. 80% of sites used in this study were classed as kerbside or roadside according to historical UK Department for Environment, Food & Rural Affairs (DEFRA) classification categories, that is, sites located within 1 and 15 m from a road, respectively.[24] Possible biases arising from design of the monitoring network are outlined in the Discussion section of this paper. GPS coordinates for site locations were checked against known locations, and corrected if positional errors were discovered (Supporting Information S1). All PDT's used in the study were supplied and analyzed by a single analytical laboratory, although analyses were conducted separately for each local authority area.

PDTs exhibit known biases relative to automatic chemiluminescence monitors[25] and DEFRA recommend "bias adjustments" to PDT NO_2 data.[24] However, as bias adjustments differed between local authorities we chose to use PDT data prior to bias adjustment. Monthly PDT concentrations were generally within 2% of simultaneous colocated automatic monitoring measurements for the Glasgow conurbation (Supporting Information S2) suggesting that our use of unadjusted data would have had limited effect on our analyses.

2.2. Buffers and Variables. Variables were classified into traffic and nontraffic categories; and circular buffers created around each monitoring site at radii of 25 to 5000 m. Buffers for traffic related variables were limited to 1000 m to be consistent with other studies.[11,18,26] Noncircular variables (e.g., 'Distance to nearest major road') were also included. Full details are provided in Supporting Information S3.

Digital road network data from the OS Mastermap data set[27] included Motorways, A roads, B roads, minor roads, and local streets. Variables representing single road classes or combinations of road classes were included, along with variables weighted by the number of lanes. Traffic counts were available for motorways and A roads,[28] but were not used as their spatial accuracy was limited.

A number of variables were examined as possible indicators of the depth and geometry of street canyons to represent the potential for reduced local dispersion of air pollutants. Building heights from light detection and ranging (LiDAR) measurements[29] were manually processed to remove nonbuilding polygons prior to calculation of variables representing building area, building volume, street configuration and visible sky. The "street configuration" variable is similar to building volume except that only buildings within buffer distances of between 20 and 50 m from the road centerline were included in the calculation to more closely represent the influence of street canyons. The methods, buffer sizes and constraints reported by Tang et al.[10] were used. The percentage of visible sky was calculated in ArcGIS 3D Analyst using the "Skyline Graph" function. We evaluated this approach against field observations and found good agreement between calculated and observed values (Supporting Information S4).

Similar themed land use classifications from the CORINE Vector Land Cover Data set for 2006[30] were combined to reduce the number of variables entered into the model. For example, land classes for green urban areas and sports and leisure were combined into a single variable. The land cover class for roads was excluded as this was included in the Mastermap data.

Population and working population in "datazones" (census output areas with 500−1,000 household residents) were obtained as midyear estimates for 2007 derived from the 2001 census.[31] Altitude above sea level (as a potential predictor of increased dispersion in higher more exposed locations) was provided as a raster layer from the Panorama digital terrain model.[32]

2.3. Model Development. The relationship between observed NO_2 and land use variables was quantified using supervised forward regression in SPSS using a similar, but not identical, approach to Gulliver et al.[18] (details in Supporting Information S5).

To estimate uncertainty introduced by the model building process, we adapted a design suggested by Gulliver et al. (2013)[18] to first stratify sites by geographical subareas (in our case local authority areas); and second to randomly split the stratified sites into four completely independent groups (numbered 1−4) each of which represented pollution variations across the geographical area of the Glasgow

DOI: 10.1021/acs.est.6b02089
Environ. Sci. Technol. 2016, 50, 11085−11093

Table 1. LUR Model Statistics for Baseline Models Developed from Three Training Groups[a]

model	included variables[b]	training "n"	training adj. R^2	LOOCV R^2	training RMSE	HO R^2	HO RMSE	HO FB
Baseline_123	11.8 + 1.25 × 10^{-5} x BUILD_VOL300 + 3.45 x10^{-3} x ABM_SUM300 − 0.037 x DIST_ABM_MIN + 1.11 x10^{-6} x ALL_URB2000	100	0.82	0.81	7	0.56	12.7	0.015
Baseline_124	9.1 + 1.704 × 10^{-5} x BUILD_VOL200 + 3.31 × 10^{-3} x ABM_MIN_SUM300 + 1.73 × 10^{-6} x ALL_URB2000 − 0.045 x DIST_ABM_MIN	100	0.73	0.7	8.8	0.82	7.9	0.003
Baseline_134	25.7 + 4.99 × 10^{-4} x BUILD_VOL500 − 0.053 x DIST_ABM_MIN − 1.44 × 10^{-6} x GREEN_RUR2000 + 2.27 × 10^{-4} x STRT_CONF20/25 + 4.99 × 10^{-4} x ABM_SUM1000 + 0.123 x MIN_SUM25	103	0.77	0.74	8.4	0.62	10	0.088
Baseline_234	9.4 + 2.35 × 10^{-5} x BUILD_VOL200 + 5.25 x10^{-3} x ABM_SUM200 + 1.60 × 10^{-6} x ALL_URB2000 − 0.04 x DIST_ABM_MIN	102	0.71	0.67	9.1	0.85	7	−0.041

[a]Variable order reflects order in which variables were entered into the model. [b]Abbreviations for included variables are as follows, and are described in further detail in section S3 of the Supporting Information. The numbers at the end of variable names refer to GIS buffer radii in meters. ABM_SUM: major road length; ABM_MIN_SUM: major and minor road length; ALL_URB: all urban area (continuous and discrete); BUILD_VOL: building volume: DIST_ABM_MIN: distance to nearest major or minor road; GREEN_RUR: green rural area; MIN_SUM: minor road length; STRT_CONF: street configuration. LOOCV: Leave one out cross validation; HO: Hold-out; RMSE: root-mean-square error; FB: Fractional bias.

conurbation. "Baseline" models, representing best case scenarios that maximized the number of training PDT sites used in model development while still maintaining approximately 25% of the total number of PDT sites as independent hold-out sites in independent groups for model evaluation, were developed from combinations of three out of the four groups. "Subset" models were developed from single groups (subsequently referred to as Subset_1G groups) or combinations of 2 groups (Subset_2G groups) to investigate the influence of the number of sites on variance explained by models. Models were named as follows: "Model_xyz" where "Model" describes the type of model, and "xyz" describes the groups included in the training data set (Supporting Information S6).

Models were evaluated against independent hold-out sites not included in model development. Where one or two groups were used in model development, models were also evaluated against a random subset (n = 33) of the hold-out sites which was resampled 500 times, and an average R^2 for the resampled hold-out data reported. This approach provides consistency to model evaluation as the sample size of the hold-out data was equal and similar to the number of hold-out sites in the baseline models. Model explained variance (Training adj. R^2), root-mean-square error (RMSE), leave one out cross validation (LOOCV R^2; for training data only) and fractional bias (FB; for hold-out data only), were calculated as metrics of model performance.

2.4. Pollution Maps and Exposure Estimation. Spatial variations in NO_2 were mapped at 10 × 10 m resolution. Details of how pollution surfaces were produced are provided in Supporting Information S7. To avoid negative concentrations during mapping, (which may have arisen because of the absence of PDT sites in rural locations in the local authority networks used for model development) the lower limit of modeled concentrations was set equal to the annual average measurement (9.7 μg m^{-3}) at the Waulkmillglen Scottish Automatic Network monitoring site located in a rural location approximately 16 km to the south of the city center. This adjustment to background concentration was required at, on average, 4.7% (min. 0%; max. 9.9%) of cohort postcode centroids (see below) depending on model involved.

Residential exposure was estimated for a cohort of 13,679 patients attending a blood pressure clinic in Glasgow City[33] using postcode centroid coordinates for each cohort address location (9,161 separate postcode centroids) to extract the modeled concentrations from the pollution surfaces. We compared exposure estimates in two ways. First, we computed within-model summary statistics for each model to characterize the distribution of residential exposures produced by the model. Second, we assessed between-model precision of exposure estimates at each cohort participant location for each type of model; that is, we calculated "intralocation" precision statistics across all models from Subset-1G, Subset-2G and Baseline models, respectively. Precision was assessed through two metrics; the difference between maximum and minimum exposure estimate at each location for each model type (subsequently referred to as "intralocation range"), and the relative standard deviation (RSD) of exposure estimates at each location for each model type (subsequently referred to as "intralocation RSD"). In addition we estimated mean exposure at each location for each model type (subsequently referred to as "intralocation mean"). Analyses were conducted with ArcGIS and R.[34,35]

3. RESULTS

3.1. PDT NO_2 Observations. Highest and lowest mean, range and maximum concentrations were observed in Glasgow City and East Renfrewshire respectively (Supporting Information S8). Glasgow City had the largest number of sites (n = 60) and largest percentage (23%) of urban background sites[24] (cf. approximately 20% background sites in amalgamated data set for the full conurbation).

The four randomly selected groups had similar distributions of NO_2 concentrations, with exception of group 2 which had a maximum concentration 30% lower than the remaining groups because of a lower proportion of very high concentration sites in group 2 (Supporting Information S9). Groups 1−4 contained 5, 9, 6, and 7 background sites and 5, 2, 5, and 5 kerbside sites, respectively.

3.2. Baseline Models. Baseline (3-group) models explained 71−82% and 56−85% of the variation in NO_2 in training and hold-out data respectively (Table 1). Models with the highest training adj. R^2 (Baseline_123 and 134) had the lowest hold-out R^2. RMSE ranged between 7−9 μg m^{-3} and 6−13 μg m^{-3} for training and hold-out sites, respectively. FB was small indicating slight over or under prediction depending on the model. LOOCV R^2 was generally 2−4% lower than Training adj. R^2 (Table 1).

DOI: 10.1021/acs.est.6b02089
Environ. Sci. Technol. 2016, 50, 11085−11093

Table 2. Model Statistics for Models Developed Using Fewer than Three Training Groups[a]

model	included variables[b]	training "n"	training adj. R^2	LOOCV R^2	training RMSE	HO R^2	HO RMSE	HO FB	mean resampled R^2 (IQR)
Subset_1G_1	$17.7 + 2.13 \times 10^{-5}$ x CONT_URB1000 + 1.58 $x10^{-6}$ x DISC_URB2000 − 0.062 x DIST_ABM_MIN + 1.51 $x10^{-3}$ x ABM_SUM400	33	0.89	0.85	5.2	0.6	11.3	0.051	0.6 (0.51−0.69)
Subset_1G_2	$5.7 + 6.61 \times 10^{-4}$ x BUILD_AREA200 + 1.89×10^{-2} x A_SUM100 + 2.44×10^{-7} x DISC_URB5000	32	0.75	0.71	7.3	0.58	11.6	0.030	0.59 (0.51−0.66)
Subset_1G_3	$22.1 + 2.29 \times 10^{-5}$ x BUIILD_VOL200 + 7.90×10^{-3} x ABM_SUM200 − 1.53x 10^{-6} x ALL_GREEN2000 + 4.00×10^{-5} x CONT_URB400	35	0.9	0.88	5.4	0.65	11.5	0.032	0.65 (0.57−0.74)
Subset_1G_4	$-0.4 + 3.12 \times 10^{-2}$ x BUILD_AREA25 + 2.20 $x10^{-3}$ x MINOR_SUM1000 + 3.65×10^{-2} x ABM_SUM100	35	0.78	0.74	8.4	0.5	12.7	0.032	0.51 (0.45−0.57)
Subset_2G_12	12.2 + 1.13×10^{-5} x BUILD_VOL300 + 3.22 $x10^{-3}$ x ABM_SUM300 + 1.26 $x10^{-6}$ x ALL_URB2000 − 0.041 x DIST_ABM_MIN	65	0.8	0.78	7.1	0.7	10	0.003	0.69 (0.61−0.79)
Subset_2G_13	24.8 + 1.29×10^{-5} x BUILD_VOL300 + 3.57 $x10^{-3}$ x ABM_SUM300 − 1.23 $x10^{-6}$ x ALL_GREEN2000 − 0.039 x DIST_ABM_MIN	68	0.86	0.84	6.3	0.61	11	0.026	0.62 (0.54−0.73)
Subset_2G_14	14.7+ 2.80×10^{-6} x BUILD_VOL500 − 0.064 x DIST_ABM_MIN + 1.79 $x10^{-6}$ x ALL_URB2000 + 3.18 x STRT_CONF_20/25	68	0.71	0.65	9.3	0.7	9.4	0.069	0.69 (0.62−0.77)
Subset_2G_23	8.6 + 3.20×10^{-5} x BUILD_VOL200 + 8.40 $x10^{-3}$ x ABM_SUM200 + 1.64 $x10^{-7}$ x ALL_URB5000	67	0.8	0.78	7.3	0.65	11.9	0.012	0.63 (0.56−0.73)
Subset_2G_34	21.5 + 1.09×10^{-4} x BUILD_AREA500 − 0.063 x DIST_ABM_MIN + 4.05 $x10^{-4}$ x STRT_CONF20/25 − 1.32 $x10^{-6}$ x GREEN_RUR2000	70	0.73	0.69	9.2	0.66	9	0.017	0.66 (0.62−0.71)
Subset_2G_24	22.6 + 1.49×10^{-4} x BUILD_AREA500 − 0.059 DIST_ABM_MIN − 1.44×10^{-6} x GREEN_RUR2000	67	0.62	0.58	10.5	0.75	10.2	−0.049	0.75 (0.72−0.79)

[a]Variable order reflects order in which variables were entered into the model. LOOCV: Leave one out cross validation; HO: Hold-out; Resampled: models were evaluated against a random subset (n = 33) of the hold-out sites which was resampled 500 times and an average R^2 reported. Parentheses indicate lower and upper quartile range respectively. [b]Abbreviations for included variables are as follows, and are described in further detail in section S3 of Supporting Information. The numbers at the end of variable names refer to GIS buffer radii in meters. ABM_SUM: major road length; ABM_MIN_SUM: major and minor road length; ALL_GREEN: all green areas (urban and rural); ALL_URB: all urban areas (continuous and discrete); A_SUM: A road length: BUILD_AREA: building area; BUILD_VOL: building volume; CONT_URB: continuous urban area; DISC_URB: discontinuous urban area; DIST_ABM_MIN: distance to nearest major or minor road; GREEN_RUR: green rural area; MIN_SUM: minor road length; STRT_CONF: street configuration. LOOCV: Leave one out cross validation; HO: Hold-out; RMSE: root-mean-square error; FB: Fractional bias.

DOI: 10.1021/acs.est.6b02089
Environ. Sci. Technol. 2016, 50, 11085−11093

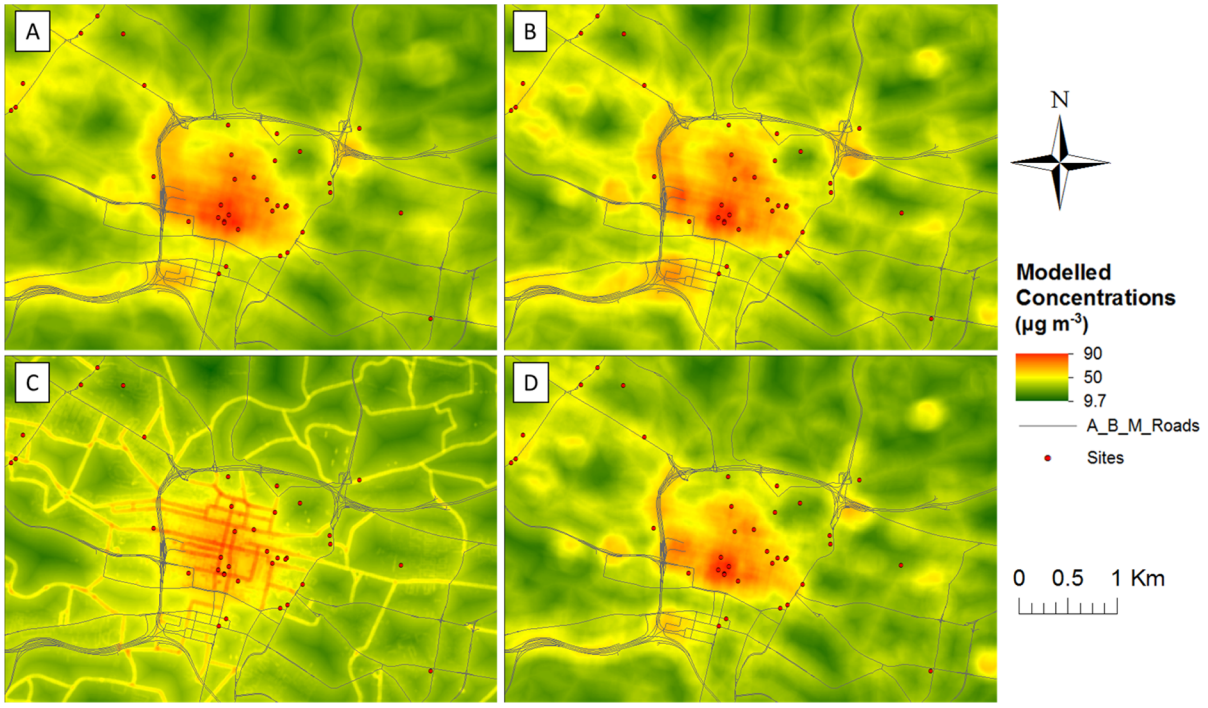


Figure 1. NO_2 pollution maps of Glasgow city center modeled by baseline models developed from three out of four randomly selected groups of passive diffusion tube monitoring sites. (A) Baseline_123, (B) Baseline_124, (C) Baseline_134, and (D) Baseline_234. Pollution maps are raster maps for each model created by summing the contribution to overall pollution estimates from each variable in the model.

Baseline model variables were relatively consistent, with all models including variables representing building volume, road length and distance to a major or minor road. Models also included buffers representing urban (Baseline_123 & Baseline_234) or green (Baseline_124 and Baseline_134) areas. Building volume explained the largest percentage of variance in all models (58−70%) followed by the sum of the length of major roads within specified buffer distances (Supporting Information S10).

Baseline models contained no influential sites (i.e., Cooks Distance[36] < 1 for all sites), but a single outlier was identified with a residual greater than 3 times the standard deviation of the residuals. Regression residuals showed no spatial autocorrelation (Morans I $p > 0.05$) (Supporting Information S10).

3.3. Sensitivity Analyses. Subset_1G models developed from 1 training group ($n \approx 33$) explained 75−89% variance in training data, 50−65% in hold-out data, and 51−65% in resampled hold-out data compared to 67−81% in training data and 51−65% in hold-out data for baseline models (Table 2). Subset_1G models showed heterogeneity in the included variables and buffer sizes and generally contained fewer variables than baseline and Subset_2G models. Training RMSE for Subset_1G models were generally lower than those observed for baseline models, while hold-out RMSE for Subset_1G models were slightly higher than those observed for baseline models (Table 2).

Subset_2G models developed from two groups ($n \approx 66$) explained similar variance in training data (62−86%) to baseline models. Explained variance in hold-out data was slightly lower for Subset_2G models compared to baseline models (hold-out explained variance R^2 61−75%; resampled hold-out explained variance R^2 62−75%; Table 2). All subset models had small FB (Table 2).

3.4. Pollution Maps and Exposure Estimation. All models predicted broadly anticipated spatial patterns in pollution with high concentrations predicted in the city center, adjacent to main roads, and around road junctions (Supporting Information S11, S12). Baseline model maps were similar for three out of four models (Figure 1), with Baseline_134 highlighting pollution contrasts around the road network more prominently than other Baseline models (Figure 1C). Median exposures of 24.9, 27.2, 26.7, and 24.9 $\mu g\ m^{-3}$ were predicted by Baseline_123, 124, 134, and 234 models respectively (Figure 2), with between 8 and 10% of the cohort group estimated to

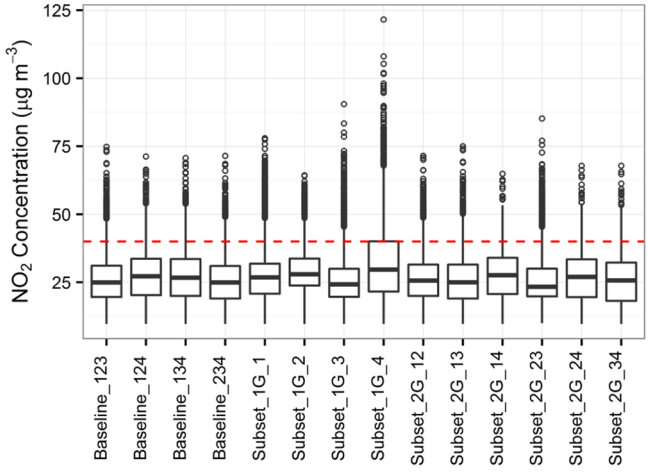


Figure 2. Boxplot of estimated exposure concentrations at cohort residential addresses for each model. Central line represents median. Box represents range. Upper and lower whiskers represent the highest and lowest datum within 1.5 times the upper and lower interquartile range, respectively. Circles reflect data outside this range. The dashed horizontal line denotes the 40 $\mu g\ m^{-3}$ UK air quality standard for NO_2.

DOI: 10.1021/acs.est.6b02089
Environ. Sci. Technol. 2016, 50, 11085−11093

be exposed to residential concentrations above the UK national air quality standard of 40 μg m^{-3} (data not shown). Baseline models estimated between 4 and 6% of the cohort were exposed to background concentrations, contributing to a bimodal distribution of mean intralocation exposure estimates (Figure 3A).

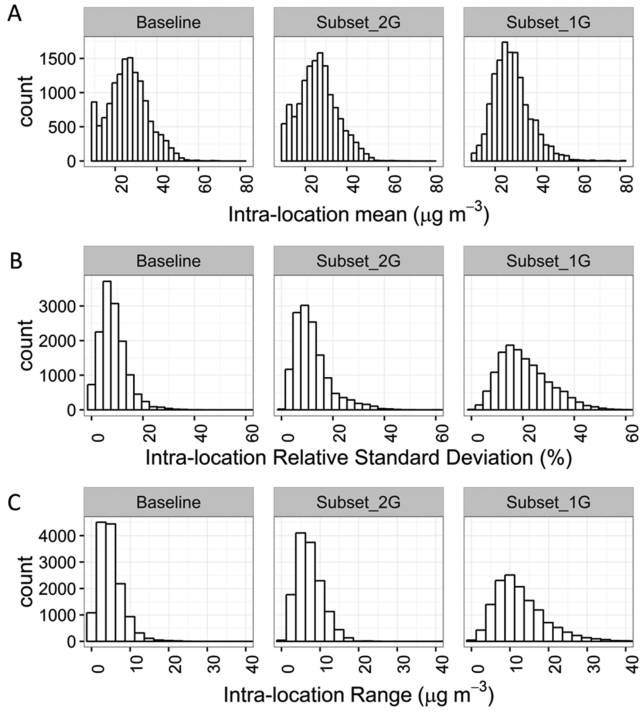


Figure 3. Exposure estimate variability for Baseline (4 models), Subset_2G (6 models), and Subset_1G models (4 models). (A) Histograms of mean intralocation predicted concentration; (B) Histograms of intralocation RSD; (C) Histograms of intralocation range of model estimates.

The distribution of cohort exposure estimates highlighted inconsistencies between Subset and Baseline models, with Subset_1G models showing the widest range of modeled concentrations (Figure 2). Subset_1G models estimated between 4% (Subset_1G_1) and 25% (Subset_1G_4) of the cohort were exposed to concentrations greater than 40 μg m^{-3}, while Subset_2G models estimated between 8 and 9% were exposed to concentrations greater than 40 μg m^{-3} (data not shown).

The median intralocation range for Subset_1G Subset_2G and Baseline models was 11.4, 6.6, and 4.3 μg m^{-3} respectively (Figure 3C). The median intralocation RSD was 19.2, 10.3, and 7.3% for Subset_1G, Subset_2G, and Baseline models respectively (Figure 3B).

4. DISCUSSION

4.1. Baseline Models. Baseline models developed from local authority PDTs explained 56−85% of the variance in NO_2 concentrations at hold-out sites; similar to model performance statistics reported for LUR models developed from purpose designed networks.[3]

Building volume was selected first in each model and explained 58−70% of variance in NO_2 concentrations in training data. Although it was unexpected that building volume would have the highest explanatory power, other studies have reported that variables other than those derived from traffic have explained the largest proportion of variance in LUR models.[11,21] The higher percentage of NO_2 variance explained by building volume compared with traffic derived variables may be a consequence of the limited traffic data used in this study (no information on traffic flow and fleet composition) as literature reported models including traffic intensity have explained, on average, an additional 10% variance compared with those which did not.[11]

An a priori assumption for building volume was that it represented the built-up area (hence human activity) in proximity to the monitoring sites. We allowed building volume and street configuration variables to coexist in the models, as studies in The Netherlands and London demonstrated improvements in explained variance in models with variables representing street configuration.[9,10] Only Baseline_134 included a street configuration variable (Table 1) and the increase in explained NO_2 variance was modest, suggesting that building volume accounted for a substantial proportion of street canyon induced concentration contrasts (Supporting Information S10). Larger buffer sizes were selected for building volume (200−500 m) compared with those selected for street configuration (25 m), suggesting that building volume represented emissions over a wider area in addition to the local influence of the street canyon. The CORINE land classification data, which also represents urban area, was included in three out of four baseline models; and whereas VIF was generally low (VIF < 2) there was potential for multicolinearity with building volume variables. This illustrates subjectivity in decisions regarding which variables should be considered mutually exclusive when establishing selection criteria prior to LUR modeling.

Similarly, assumptions about anticipated direction of associations were not always clear. This study frequently found "local street" variables to be significant in model development; however these variables had a negative coefficient implying increased local streets reduce NO_2 concentrations. Thus, in relation to the anticipated sign of effect, local streets were removed during model development. Local streets tended to be located away from busier roads (and hence elevated pollution concentrations) and therefore this potentially valid, though counterintuitive, association was omitted from the models reported here.

Separating the data into groups and development of baseline models from three out of four groups, had a modest influence on training and hold-out explained variance, and variable selection. Site residuals also showed consistency in direction and magnitude between models, irrespective of the groups used to build the models (Supporting Information S14). The lower explained variance observed on hold-out data for the "best" training models (and vice versa) may imply that models are strongly influenced by a relatively small number of sites and the extent to which those sites are characterized by the GIS variables in the study. The outlying site provides an extreme example of this. This site was located in the city center in a pedestrian square away from direct emission sources but near to a very large shopping center. Measured concentrations were low, but due to the influence of the building volume variable, modeled concentrations were high. Removing this site from the hold-out data in Baseline_123 increased explained variance from 56% to 72%.

4.2. Sensitivity Analyses. Subset_1G models ($n \approx 33$) developed from independent PDT networks that did not share

DOI: 10.1021/acs.est.6b02089
Environ. Sci. Technol. 2016, 50, 11085−11093

any training sites had heterogeneous selected variables between models. Only Subset_1G_3 included building volume contrasting with Baseline models where all models selected building volume as the first variable. Subset_1G models had substantially higher training adj. R^2, and lower hold-out R^2 (Table 2) than other models, highlighting a risk of overfitting in models developed from small numbers of training sites and large numbers of variables, consistent with findings of Wang et al.,[15] Johnson et al.,[14] and Basagaña et al.[13]

Subset_2G models ($n \approx 66$ sites) explained, on average, 9.5% more, and 3.5% less variance in hold-out data compared to Subset_1G and Baseline models, respectively. Remaining model statistics, and included predictors, were similar between Subset_2G and Baseline models. These findings, and those of Basagaña et al., who concluded that >80 sites are desirable for LUR model development, highlight a risk of misclassification of exposure from models developed from fewer sites.

4.3. Exposure Estimation. Subset_1G models involved a similar number of training sites to some LUR studies in the literature,[3] and their development from independent groups of PDT sites provided an opportunity to examine the potential variability of cohort exposure estimates developed from models developed from small networks of sites. The influences of training site number on model performance has been reported in other studies.[15,37,38] However, these studies compared models developed using different sampling regimes, timeframes, and spatial scales; and did not examine distributions of cohort exposure estimates. Consequently these earlier studies are relevant to, but not directly comparable with our analyses.

Subset_1G models generated pollution maps with marked differences in spatial patterns between models (Supporting Information S12 and S13). Model differences are reflected in exposure estimates at cohort locations; with median intralocation range and RSD of 11.4 μg m^{-3} and 19.2% respectively across the four models; and 4−25% of the cohort group estimated to be exposed to concentrations greater than 40 μg m^{-3} depending on the group of sites used to develop the model. This illustrates that potential for misclassification of exposure between Subset_1G models is markedly greater than for the Baseline and Subset_2G models.

Variations in Subset_1G exposure estimates result, in part, from the wider range of estimates from the Subset_1G_4 model. Subset_1G_4 explained the lowest percentage of variance in hold-out data (Table 2), although this alone does not explain the higher variability in exposure estimates for Subset_1G models as a whole, as removal of Subset_1G_4 only reduced the median intralocation range and RSD to 7.6 μg m^{-3} and 14.9% respectively (i.e., variance of restricted set of Subset_1G models remained higher than Subset_2G models despite removal of Subset_1G_4).

Variations in intralocation estimates between models are unlikely to result from the number of background and kerbside sites, or the distribution of NO_2 concentrations within each group, as these were broadly similar between groups (Supporting Information S9). Furthermore, Subset_1G_2, which contained the lowest range of PDT concentrations and the lowest maximum concentration, did not give substantially different exposure estimates at the cohort locations (Figure 2). Variations in Subset_1G exposure estimates may instead reflect inherent variations which might be expected if LUR models with similar numbers of sites are redeveloped for the same geographic area using alternative PDT networks, particularly if large numbers of variables are used in model development.[13] However, it is also possible that our study overestimated variability between Subset_1G models. Many LUR studies expend considerable effort selecting sites in sampling networks.[39] In contrast our study generated four random networks as a subset of a larger network, which was comprised of four independent local authority PDT networks.

Subset_2G models ($n \approx 66$) showed increased precision compared to Subset_1G models, with median intralocation range and RSD of 6.7 μg m^{-3} and 10.3% respectively. Subset_2G models estimated that between 8 and 9% of cohort participants were exposed to residential NO_2 concentrations >40 μg m^3, which was similar to the estimates provided by Baseline models supporting the suggestion that a minimum of 60 sites are required for reliable LUR estimates.

4.4. Limitations. Biases may have been introduced into our LUR models through characteristics of the local authority PDT networks used. 80% of PDT sites were at kerbside and roadside locations where concentrations are anticipated to approach or exceed air quality standards, introducing possible biases from "preferential sampling".[17] This proportion of kerbside and roadside sites is greater than many LUR models reported in the literature where typically 50−75% of monitoring sites are in background locations.[15,21,40,41]

Entirely unbiased comparisons between 1, 2, and 3-group models could not be made. Subset_1G models did not share training sites and so were fully independent of each other. In contrast models developed from 2 and 3 groups shared up to 50 and 66% respectively of the training sites between models (e.g., Subset_2G_12 shared 50% of the training sites with Subset_2G_14). This may have constrained the selection of predictors, and may consequently have improved the precision of Baseline and Subset_2G exposure estimates. A PDT network in excess of 400 sites would be required to obtain four fully independent replicates of Baseline models.

With appropriate acknowledgment of the above limitations, our analyses allow estimation of the effect of use of different sizes of network on LUR model development and exposure estimates within realistic constraints of available pollution monitoring networks. An "ensemble" exposure prediction of the average of the four baseline models would be an appropriate way to apply these model estimates in epidemiological analyses.[18]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.est.6b02089.

Additional information including summary statistics for PDT concentrations, evaluation of PDT against reference analysers, baseline model statistics and residuals, and modeled pollution surfaces (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +44 141-548-3202; e-mail: Iain.Beverland@strath.ac.uk.

Author Contributions

J.G. and I.B. designed the study. J.G. collated data, developed and evaluated models, and conducted data analyses. The first draft of the manuscript was manuscript was written by J.G. and all other authors contributed to discussions on data analysis and

DOI: 10.1021/acs.est.6b02089
Environ. Sci. Technol. 2016, 50, 11085−11093

revisions of the paper. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

The research data associated with this paper are available at: http://dx.doi.org/10.15129/1692520b-deb7-4571-a401-3a870fe52a31.

■ ACKNOWLEDGMENTS

We thank Dr Patricia Drach and Dr Rohinton Emmanuel (Glasgow Caledonian University) for providing measurements of visible sky in the Glasgow conurbation which were used to evaluate ArcGIS functions. J.G. is funded jointly through an Engineering and Physical Sciences Research Council Doctoral Training Grant (EPSRC DTG EP/L505080/1 and EP/K503174/1) studentship, supervised by Iain Beverland and Scott Hamilton, with support from the University of Strathclyde and Ricardo Energy and Environment.

■ ABBREVIATIONS

land use regression	LUR
passive diffusion tube	PDT
nitrogen dioxide	NO_2
geographic information system	GIS
leave one out cross validation	LOOCV

■ REFERENCES

(1) WHO. Review of evidence on health aspects of air pollution − REVIHAAP Project http://www.euro.who.int/en/what-we-do/health-topics/environment-and-health/air-quality/publications/2013/review-of-evidence-on-health-aspects-of-air-pollution-revihaap-project-final-technical-report (accessed September 16, 2013).

(2) BRIGGS, D. J.; COLLINS, S.; ELLIOTT, P.; FISCHER, P.; KINGHAM, S.; LEBRET, E.; PRYL, K.; VAN REEUWIJK, H.; SMALLBONE, K.; VAN DER VEEN, A. Mapping urban air pollution using GIS: a regression-based approach. *Int. J. Geogr. Inf. Sci.* **1997**, *11* (7), 699−718.

(3) Hoek, G.; Beelen, R.; de Hoogh, K.; Vienneau, D.; Gulliver, J.; Fischer, P.; Briggs, D. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmos. Environ.* **2008**, *42* (33), 7561−7578.

(4) Arain, M. A.; Blair, R.; Finkelstein, N.; Brook, J. R.; Sahsuvaroglu, T.; Beckerman, B.; Zhang, L.; Jerrett, M. The use of wind fields in a land use regression model to predict air pollution concentrations for health exposure studies. *Atmos. Environ.* **2007**, *41* (16), 3453−3464.

(5) Mavko, M. E.; Tang, B.; George, L. A. A sub-neighborhood scale land use regression model for predicting NO2. *Sci. Total Environ.* **2008**, *398*, 68−75.

(6) Li, X.; Liu, W.; Chen, Z.; Zeng, G.; Hu, C. M.; León, T.; Liang, J.; Huang, G.; Gao, Z.; Li, Z.; et al. The application of semicircular-buffer-based land use regression models incorporating wind direction in predicting quarterly NO2 and PM10 concentrations. *Atmos. Environ.* **2015**, *103*, 18−24.

(7) Dons, E.; Van Poppel, M.; Kochan, B.; Wets, G.; Int Panis, L. Modeling temporal and spatial variability of traffic-related air pollution: Hourly land use regression models for black carbon. *Atmos. Environ.* **2013**, *74*, 237−246.

(8) Mölter, A.; Lindley, S.; de Vocht, F.; Simpson, A.; Agius, R. Modelling air pollution for epidemiologic research − Part II: Predicting temporal variation through land use regression. *Sci. Total Environ.* **2010**, *409* (1), 211−217.

(9) Eeftens, M.; Beekhuizen, J.; Beelen, R.; Wang, M.; Vermeulen, R.; Brunekreef, B.; Huss, A.; Hoek, G. Quantifying urban street configuration for improvements in air pollution models. *Atmos. Environ.* **2013**, *72*, 1−9.

(10) Tang, R.; Blangiardo, M.; Gulliver, J. Using Building Heights and Street Configuration to Enhance Intraurban PM10, NOX, and NO2 Land Use Regression Models. *Environ. Sci. Technol.* **2013**, *47* (20), 11643−11650.

(11) Beelen, R.; Hoek, G.; Vienneau, D.; Eeftens, M.; Dimakopoulou, K.; Pedeli, X.; Tsai, M.-Y.; Künzli, N.; Schikowski, T.; Marcon, A.; et al. Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe − The ESCAPE project. *Atmos. Environ.* **2013**, *72*, 10−23.

(12) Eeftens, M.; Beelen, R.; de Hoogh, K.; Bellander, T.; Cesaroni, G.; Cirach, M.; Declercq, C.; Dėdelė, A.; Dons, E.; de Nazelle, A.; et al. Development of Land Use Regression Models for PM2.5, PM2.5 Absorbance, PM10 and PMcoarse in 20 European Study Areas; Results of the ESCAPE Project. *Environ. Sci. Technol.* **2012**, *46* (20), 11195−11205.

(13) Basagaña, X.; Rivera, M.; Aguilera, I.; Agis, D.; Bouso, L.; Elosua, R.; Foraster, M.; de Nazelle, A.; Nieuwenhuijsen, M.; Vila, J.; et al. Effect of the number of measurement sites on land use regression models in estimating local air pollution. *Atmos. Environ.* **2012**, *54*, 634−642.

(14) Johnson, M.; Isakov, V.; Touma, J. S.; Mukerjee, S.; Özkaynak, H. Evaluation of land-use regression models used to predict air quality concentrations in an urban area. *Atmos. Environ.* **2010**, *44* (30), 3660−3668.

(15) Wang, M.; Beelen, R.; Eeftens, M.; Meliefste, K.; Hoek, G.; Brunekreef, B. Systematic Evaluation of Land Use Regression Models for NO2. *Environ. Sci. Technol.* **2012**, *46*, 4481−4489.

(16) Babyak, M. A. What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. *Psychosom. Med.* **2004**, *66* (3), 411−421.

(17) Lee, D.; Ferguson, C.; Scott, E. M. Constructing representative air quality indicators with measures of uncertainty. *J. R. Stat. Soc. Ser. A Stat. Soc.* **2011**, *174* (1), 109−126.

(18) Gulliver, J.; de Hoogh, K.; Hansell, A.; Vienneau, D. Development and Back-Extrapolation of NO2 Land Use Regression Models for Historic Exposure Assessment in Great Britain. *Environ. Sci. Technol.* **2013**, *47* (14), 7804−7811.

(19) Chen, L.; Wang, Y.; Li, P.; Ji, Y.; Kong, S.; Li, Z.; Bai, Z. A land use regression model incorporating data on industrial point source pollution. *J. Environ. Sci.* **2012**, *24* (7), 1251−1258.

(20) Amini, H.; Taghavi-Shahri, S. M.; Henderson, S. B.; Naddafi, K.; Nabizadeh, R.; Yunesian, M. Land use regression models to estimate the annual and seasonal spatial variability of sulfur dioxide and particulate matter in Tehran, Iran. *Sci. Total Environ.* **2014**, *488−489*, 343−353.

(21) Vienneau, D.; de Hoogh, K.; Beelen, R.; Fischer, P.; Hoek, G.; Briggs, D. Comparison of land-use regression models between Great Britain and the Netherlands. *Atmos. Environ.* **2010**, *44* (5), 688−696.

(22) Gulliver, J.; Morris, C.; Lee, K.; Vienneau, D.; Briggs, D.; Hansell, A. Land Use Regression Modeling To Estimate Historic (1962−1991) Concentrations of Black Smoke and Sulfur Dioxide for Great Britain. *Environ. Sci. Technol.* **2011**, *45*, 3526−3532.

(23) LAQM reports - Air Quality in Scotland http://www.scottishairquality.co.uk/news/reports?view=laqm (accessed June 23, 2015).

(24) DEFRA. Local air quality management: Technical guidance LAQM.TG(09) - Publications - GOV.UK https://www.gov.uk/government/publications/local-air-quality-management-technical-guidance-laqm-tg-09 (accessed December 2, 2014).

(25) Heal, M. R.; O'Donoghue, M. A.; Cape, J. N. Overestimation of urban nitrogen dioxide by passive diffusion tubes: a comparative exposure and model study. *Atmos. Environ.* **1999**, *33* (4), 513−524.

(26) Kashima, S.; Yorifuji, T.; Tsuda, T.; Doi, H. Application of land use regression to regulatory air quality data in Japan. *Sci. Total Environ.* **2009**, *407* (8), 3055−3062.

(27) OS MasterMap ITN Layer [GML geospatial data], Coverage: Glasgow, Updated: Oct 2013 Year, Ordnance Survey (GB), Downloaded: October 2013 Using EDINA Digimap Ordnance Survey Service http://edina.ac.uk/digimap (accessed August 29, 2014).

DOI: 10.1021/acs.est.6b02089
Environ. Sci. Technol. 2016, 50, 11085−11093

(28) Department for Transport: Transport Statistics. http://www.dft.gov.uk/traffic-counts/.
(29) NERC Earth Observation Data Centre. Landmap; The GeoInformation Group (2014): UK building heights http://catalogue.ceda.ac.uk/uuid/e1449eac35108b49d3a0af26bb6d2060.
(30) Corine Land Cover 2006 seamless vector data - version 16 (04/2012) http://www.eea.europa.eu/data-and-maps/data/clc-2006-vector-data-version-2.
(31) Scottish Neighbourhood Statistics, Information about Scotland's Areas http://www.sns.gov.uk/ (accessed August 5, 2012).
(32) Ordnance-Survey. PANORAMA DTM data set. 1993. http://edina.ac.uk/digimap/ (accessed 2006).
(33) Aubinière-Robb, L.; Jeemon, P.; Hastie, C. E.; Patel, R. K.; McCallum, L.; Morrison, D.; Walters, M.; Dawson, J.; Sloan, W.; Muir, S.; et al. Blood pressure response to patterns of weather fluctuations and effect on mortality. *Hypertension* **2013**, *62* (1), 190−196.
(34) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2012.
(35) Hijmans, R. raster: Geographic Data Analysis and Modeling. R package version 2.4-15. http://CRAN.R-project.org/package=raster.
(36) Heiberger, R. M.; Holland, B. *Statistical Analysis and Data Display: An Intermediate Course with Examples in S-plus, R, and SAS*; Springer, 2010.
(37) de Nazelle, A.; Aguilera, I.; Nieuwenhuijsen, M.; Beelen, R.; Cirach, M.; Hoek, G.; de Hoogh, K.; Sunyer, J.; Targa, J.; Brunekreef, B.; et al. Comparison of performance of land use regression models derived for Catalunya, Spain. *Atmos. Environ.* **2013**, *77*, 598−606.
(38) Dijkema, M. B.; Gehring, U.; van Strien, R. T.; van der Zee, S. C.; Fischer, P.; Hoek, G.; Brunekreef, B. A Comparison of Different Approaches to Estimate Small-Scale Spatial Variation in Outdoor NO2 Concentrations. *Environ. Health Perspect.* **2010**, *119* (5), 670−675.
(39) Brunekreef, B. ESCAPE Project - Study Manual http://www.escapeproject.eu/manuals/ESCAPE-Study-manual_x007E_final.pdf (accessed September 24, 2015).
(40) Beelen, R. M. J.; Voogt, M.; Duyzer, J.; Zandveld, P.; Hoek, G. Comparison of the performances of land use regression modelling and dispersion modelling in estimating small-scale variations in long-term air pollution concentrations in a Dutch urban area. *Atmos. Environ.* **2010**, *44* (36), 4614−4621.
(41) Hochadel, M.; Heinrich, J.; Gehring, U.; Morgenstern, V.; Kuhlbusch, T.; Link, E.; Wichmann, H.-E.; Krämer, U. Predicting long-term average concentrations of traffic-related air pollutants using GIS-based information. *Atmos. Environ.* **2006**, *40* (3), 542−553.

DOI: 10.1021/acs.est.6b02089
Environ. Sci. Technol. 2016, 50, 11085−11093